Automatically analysing large texts in a GIS environment: The Registrar General's reports and cholera in the nineteenth century

Patricia Murrieta-Flores[1], Alistair Baron[2], Ian Gregory[3], Andrew Hardie[4] and Paul Rayson[2]

[1]University of Chester, History and Archaeology Department
[2]Lancaster University, School of Computing and Communications
[3]Lancaster University, History Department
[4]Lancaster University, Department of Linguistics and English Language

Corresponding author: Patricia Murrieta-Flores; p.murrietaflores@chester.ac.uk

Short Title: Automatically analysing large historical texts in a GIS environment

Keywords: Geographic Information Systems, Natural Language Processing, Corpus Linguistics, Spatial analysis, Humanities, Historical GIS, Digital Humanities.

Abstract

The aim of this article is to present new research showcasing how Geographic Information Systems in combination with Natural Language Processing and Corpus Linguistics methods can offer innovative venues of research to analyse large textual collections in the Humanities, and particularly, in historical research. Using as example parts of the collection of the Registrar General's Reports that contain more than 200,000 pages of descriptions, census data and vital statistics for the UK, we introduce newly developed automated textual tools and well known spatial analyses used in combination to investigate a case study about the references made to cholera and other diseases in these historical sources, and their relationship to place-names during Victorian times. The integration of such techniques has allowed us to explore in an automatic way this historical source containing millions of words, to examine the geographies depicted in it, and to identify textual and geographic patterns in the corpus.

1. Introduction

Research topics in Humanities are often concerned with geographies. Including understanding the role that landscape plays in symbolic representations depicted in material culture, the impact that spatial relationships have in the development of historical economies, or the narrative construction of places in literary works, humanities-based disciplines such as Archaeology, History and Literature have explored at different rhythms and scales issues of place and space. The approaches to 'spatiality' taken by these disciplines have been diverse and dependent on the development of their own theoretical currents, as well as to the integration of interdisciplinary research to their fields. The advent of Geographic Information Systems (GIS) was particularly welcomed and the technology promptly adopted from its emergence in those areas that deal often with spatially related quantitative data such as Archaeology (Wheatley and Gillings, 2002; Conolly and Lake, 2006). Over the last decade, crucial developments in what has been called "Historical GIS" have also changed the way historians are approaching and using quantitative and qualitative sources to advance knowledge in environmental, economic and demographic history (Gregory, 2003; Gregory and Healy, 2007; Knowles, 2008). In the case of Literature, it has been only recently that the discipline opened new venues of research, looking to explore the spatial nature of texts, attending to the so-called 'spatial turn' (Moretti, 2005; 2013; Cooper and Gregory, 2011; Gregory and Hardie, 2011; On the spatial turn see - Darby, 1953; Wright, 2005; Warf and Arias, 2008). The essential core of many such disciplines is written material and therefore, in order to take further the exploration of the geographic and spatial dimensions of diverse subjects within Humanities, a crucial step is to be able to integrate and analyse textual sources with spatial technologies. Moreover, with the creation of digital libraries, and vast digitisation efforts involving libraries, publishers and commercial organisations, new possibilities have arisen allowing the researcher to browse, analyse and query large text collections.

Two main problems have been identified in humanities research concerning collections containing large number of words (Moretti, 2005; 2013; Gregory and Hardie, 2011). The first problem is that

close reading (the traditional approach to analysing texts in History and Literature) can be too time consuming. Tools and methodologies from other disciplines that have long been text-focussed can be employed to address this issue. Natural Language Processing (NLP), a sub-discipline of computer science, targets the automatic analysis of human language by computer. NLP tools and techniques are also used in Linguistics for the study of language, and a sub-discipline has emerged over the last 40 years called corpus linguistics (CL) (McEnery and Hardie, 2012), where large bodies of naturally occurring text (corpora) are used as source material for linguistic investigations. CL and NLP methods are often used on vast quantities of text, of the order of billions of running words, and hence are able to scale to the large collections now emerging in humanities areas. The second problem is that due to the size that such collections can reach, the exploration of the overall geographies depicted within them can also prove challenging. Because of this, an increasing number of scholars are turning to the possible use of technologies, not only for the analysis of these collections, but also to examine the narratives within in terms of their spatiality. So far, research concerning corpora that contain geographic references has primarily involved the geoparsing of documents, which consists in the identification of place-names in a text, assigning geographic identifiers (coordinates) to them, and producing from these diverse visualisations such as map-based interfaces. Although essential, this effort only represents the first step to explore their geographical nature, but some exploratory studies have already shown great promise (Hardie, McEnery and Piao, 2010; Rayson, Baron and Hardie, 2012; Murrieta-Flores, Cooper and Gregory, 2012; Murrieta-Flores et al, 2013; Rupp, Donaldson and Murrieta, 2013). The identification and mapping of place-names, allows a general examination of the geographies within the texts. However, going a step further using GIS and spatial analysis will not only facilitate the identification of possible interesting patterns, but it will also advance the enquiries from these disciplines, opening new ways of understanding and interpreting the texts.

This article presents research combining NLP, CL and GIS techniques, and demonstrates how the adaptation of linguistic and spatial methods can help in the analysis of large corpora. More specifically, we present a newly developed tool (Geographical Collocates Tool), which by means of NLP techniques allows the automated extraction from a sample corpus of place-names occurring in proximity to words associated with a specific theme of interest (particular diseases in this case). This enables the automatic production of GIS-ready data files drawn from the place-names identified. In a second phase, spatial analysis methodologies (Besag-Newell method, Kulldorff spatial scan method and theSimilarity index test) were applied to the texts, leading to a detailed understanding of their spatial content. In other words, we have used automated and semi-automated approaches in order to: (1) enable the extraction of place-names related to specific topics of interest and their textual context in large corpora; (2) produce an extended overview of the geographies depicted in texts through mapping; and (3) adapt spatial analysis techniques for the detection of patterns, thus helping to point out sections of the text that call for a more extensive reading.

To demonstrate the methodologies developed, our approach has been applied to the Histpop corpus (http://www.histpop.org). This includes the digitised volumes of the U.K. Registrar General's reports, containing statistics on births, deaths and marriages from England, Scotland and Wales produced from 1801 to 1937, along with detailed textual descriptions of how the Registrar General interpreted these statistics. A case study analysing the geographical distribution of the citation of place-names associated with cholera and related diseases in the textual descriptions of the reports from England and Wales is presented for the period from the 1840s to the 1880s. This is the period in which the work of John Snow, the physician that established the contagion mode of cholera and is considered one of the fathers of modern epidemiology, did much to develop our understanding of diseases and their transmission, as well as the use of spatial thinking for studying contagion.

2. NLP, spatial analysis and the analysis of texts

The fields of Corpus Linguistics and NLP are concerned with the computer-aided analysis of large bodies of texts. Corpus linguistic methods can assist in the investigation of the usage of the language in a particular corpus, exploring lexis, grammar, semantics, and discourse structure among many other related issues (McEnery, Xiao and Tono, 2006; McEnery and Hardie, 2012). In their most basic form, corpus analyses provide frequency counts of items encountered in the text such as words and phrases. In the particular context of this research, such an approach becomes relevant as it enables us not only to search for and identify place-name instances within the corpus (from which results can be mapped), but also their *concordances*: the words that occur near the place-name, providing it a context. In addition, *Collocation Analysis* makes possible the search and extraction of terms within the texts that are associated (or that 'collocate') with any other particular word. This allows us to search within a boundary, such as a paragraph, for place-names within the corpus that are associated with themes or topics of interest. An example of such a query could be place-names collocating in the text with the words "flu" or "small-pox" (Fig. 1). Place-names in the text could be marked up manually, but if we use NLP techniques of named-entity recognition (NER - the process of locating, classifying and annotating named elements, such as people, organisations or places, in running text), it is possible to extract place-names in the text automatically to a certain degree of accuracy. In this manner, NLP provides the tools not only to quantify the number of times a particular place is mentioned - giving a possible measure of the attention it received in the texts - but when linked with collocation analysis, to identify potential topics to which that place might be related. As it is feasible to access in an efficient way the context in which places are mentioned, qualitative research can also be carried out.

Our project developed for this research the online ' 'Geographical Collocates Tool' (GCT) (a screenshot of the interface is given in Fig. 3) that enables collocation analysis of place-names and the automatic formatting of the resulting data in GIS-ready files. Given a corpus of texts in which mentions of place-names have been annotated with XML tags, the GCT can be used to search for

specific words or topics; it returns a list of place-name mentions that occur in the textual proximity of the searchedfor word or topic. The GCT relies on the corpus already having place-names marked up. This can be performed manually, as in the case of the smaller Lake District corpus used in our project (Murrieta-Flores et al, 2013); or automatically using NER tools, as described in section 3 for the larger Histpop collection. Elsewhere (Rupp et al, 2013), we describe the manual annotation of place-names and how we are customising an existing automatic NER system for historical texts. CQPweb, a web-based corpus analysis system described in detail by Hardie (2012), is used as the basis of GCT's retrieval functionality. When a corpus is loaded into CQPweb, a quickly-searchable index is built of every word contained within it. However, for our purposes we *also* index various layers of word-level linguistic annotation such as part-of-speech tags and semantic categories. These annotations are added beforehand, usually by automatic means, using e.g. CLAWS (*Constituent Likelihood Automatic Word Tagging System*: Garside and Smith, 1997) for part-of-speech tagging and USAS (*UCREL Semantic Analysis System*: Rayson et al, 2004) for semantic category tagging; the detailed workings of corpus annotation software are beyond the scope of this paper, but like much such software, CLAWS and USAS assign one or more category labels (grammatical and semantic respectively) to each word token in their input data using a combination of predefined linguistic knowledge resources such as lexicons, and probabilistic contextual disambiguation. Every multiword structure in the input data is also indexed, e.g. sentences, paragraph, and place-name mentions, working from a simple XML-based representation in the index data. All indexed features of the input corpus are then available for query and retrieval. During pre-processing, the GCT queries CQPweb to search for all annotated place-name mentions. For each mention found, it stores in a separate relational database its position in the corpus, alongside any associated spatial metadata (e.g. longitude, latitude, place type, etc.) created during georeferencing and present in the XML attributes, and the boundary positions of the structures it is contained within (e.g. sentence, paragraph and text).

To perform a search with the GCT, the user first defines the search query; this can be a single word, a sequence of words, a list of alternative words, a linguistic annotation (e.g. a semantic category), or any combination of the above. CQPweb is used to perform the search, resulting in a list of 'hits' from the corpus. From this list, the corpus position of each hit and the structures (e.g. sentence, paragraph) within which the hit occurs can be extracted. The user also provides a proximity to the search term in which to look for place-name mentions. This proximity might be a number of words (e.g. 5 words left or 5 words right), in which case the index of each hit is compared to the index of each place-name mention, if the indexes are within the specified number of words the place-name mention is marked as one that occurs within the proximity of the search term. The proximity could also be defined as within the same structure, e.g. same sentence or same paragraph, in which case the index of the structure containing each hit is compared to the index of the structure containing each place-name mention, and any matches are marked as within proximity of the search term. The two proximity options can also be combined, so for example, a query such as "look 5 words left and right, as long as it is also within the same sentence" could be constructed. Relational database joins are used to perform these search queries. If metadata about each text is also available, this is also stored in the GCT's database, meaning that search queries can also define a filter so that place-name mentions are only returned from specific texts (e.g. from certain time periods). It is also possible to filter based on place-name metadata, so only place-names of a certain type (e.g. populated places) or within longitude and latitude boundaries are returned.

Of course, there is no guarantee that, in every case where a hit is made for a place-name in proximity to a queried term or concept, an *explicit* link is made in the text between that place and that term or concept. For example, imagine we have run a query for the semantic category of *War*, with a proximity span of 5 words left, 5 words right. The GCT system would detect the place-name *London* in both of the following (hypothetical) sentences: (i) *Before he fought in the Boer War, Fred lived in London for five years*. (ii) *The army approached London*. In (ii), the term tagged having *War*

(i.e. *army*) has a grammatical link to *London*: *army* is the subject of the verb of which *London* is the object, and moreover subject, verb and object together encode a single state-of-affair in which London is clearly involved and which clearly has to do with War. There is, then, an explicit link between place and concept. In (i), none of the same factors are present: the *War* term is close to *London*, but there is no single state-of-affairs that links them, and thus no explicit connection is made. One might question whether the fact that the GCT will identify cases similar to (i) as readily as cases similar to (ii) invalidates its output. There are at least two reasons to believe that this is not the case. First, work in linguistics on collocation – a field to voluminous to review in any detail here, but see for instance Sinclair (1991), Sinclair et al. (2004) – seems to suggest that simple proximity, with no explicit connection, if repeated consistently, is indeed enough to create an implicit link between two meanings in the mind of speakers and hearers (see especially Hoey 2005). Second, even if some hits to the query do represent random juxtaposition – which we must expect to be the case at least part of the time – we would expect the types of juxtaposed element to be randomly-distributed. This adds noise to the analysis, but typically not to a sufficient degree to drown out signal the repeated, consistent co-occurrence patterns which are the focus of the qualitative analysis of the mapped form of the data that the GCT ultimately allows the researcher to generate.

GCT thus produces tables of place-name mentions found within a defined proximity of a given search term. Contained in these tables are: all spatial metadata relating to the place-name (e.g. name, type, longitude and latitude - whatever is contained as XML attributes of the tagged place-name), the surrounding context (running text to the left and right) of both the search term and the place-name mention, and any metadata present relating to the source text (e.g. date, text type, etc.). Other tables can also be produced, including a list of all place-names found with counts for each of how many times it is mentioned in the whole corpus and in texts grouped by metadata (e.g. decade, text-type, author, etc.). Once a geo-dataset has been created from the GCT output, the next

stage is to map the place-names found (using the spatial metadata) and to carry out any GIS-based analysis desired.

The discovery and understanding of spatial patterns have been at the core of GIS since its conception. Diverse methods and techniques have been developed in several scientific disciplines to uncover spatial patterns related to phenomena such as crime and epidemics. In the Humanities, spatial analysis has been used extensively in Archaeology and more recently in History (Gregory, 2003; Wheatley and Gillings, 2004). Methodologies such as cluster detection, hot spot analysis and regression models, originally developed for environmental management and urban planning, have been successfully adapted and used within Humanities research for some time (Lock, 2000; Gregory, Marti-Henneberg and Tapiador, 2010; Chrysanthi, Murrieta-Flores and Papadopolous, 2012). Nevertheless, these approaches have mainly been used on quantitative sources rather than applied to the study of texts, and as large corpora become available in digital form, there is a pressing need to develop approaches that allow these to be analysed. In this vein, this paper illustrates how methodologies developed for point pattern analysis can be used in the examination of the places depicted in large volumes of text, investigating not only where the corpus is talking about but also the geographies associated with particular themes, in this case a set of diseases.

Cluster detection methods such as the Besag-Newell technique or Kulldorff's spatial scan statistic have been widely implemented in bio-medical research, and they are among the most popular methods used to analyse whether diseases such as leukemia, diabetes, and malaria amongst others are clustered in space and/or time (Besag and Newell, 1991; D'Aignaux et al. 2002; Green et al, 2003; Brooker et al, 2003; Zhang et al, 2008; Chen et al, 2008). These techniques are well known, and they can be found in programs such as Clusterseer (http://www.biomedware.com), SaTScan (http://www.satscan.org) and R (http://www.r-project.org/) . In this paper both methods are used to explore the spatial distribution of place-name references in the Histpop corpus and their association,

as seen in the next section, to a topic of our interest. The idea behind this was to test their capacity to identify patterns in the discourse related to the geographies depicted within the historical documents.

In addition, this study also employed the analysis called Similarity index (S-Index) to assist with investigating spatial patterns (Andresen, 2009; Andresen and Malleson, 2011). The S-Index calculates the degree of similarity between two different point distributions occurring in the same area, and it allows us to map where the spatial pattern differs (Andresen, 2009). This approach has been previously used to investigate crime comparing the spatial distributions of diverse types of offence, and the test is available through a standalone application called SpatialTest (https://code.google.com/p/spatialtest/) (Andresen, 2009; Andresen and Malleson, 2011). In our case, this technique allowed us to examine further hypotheses to investigate the texts, analysing changes over time on the spatial information conveyed in the historical reports.

Although, as noted above, these spatial analysis techniques have been applied in a variety of disciplines, they have never been used for the analysis of written sources and we contend that they can be used to explore the geographies described within large volumes of geo-specific texts. Furthermore, we contend that they can help not only to solve particular research questions, but also to reveal, identify and visualise patterns that traditional research methods either struggle or fail to detect.

3. Data and processing: Reporting disease in England and Wales during Victorian times

To illustrate and evaluate our hybrid NLP-CL-GIS methodology, in this article we present a case study showing the application to a large collection of historical material. The Histpop collection is composed of more than 200,000 pages of census and registration material from the British Isles. In this case, the entire collection was geoparsed by Claire Grover and her colleagues at the University

of Edinburgh (see Grover et al, 2010 for a full description of the geoparsing process). Geoparsing poses diverse challenges and, although this will not be explored in this article, our project is also currently working on the expansion of gazetteers and improvement of NER techniques and georeferencing of different geoparsing tools (including the Edinburgh Geoparser), especially when applied to historical texts (Rayson, Baron & Hardie, 2012; Rupp et al, 2013).

In the case of the Registrar General's reports, approximately 73,322 place-names were identified by the Edinburgh Geoparser amongst the 5.4 million words that correspond to these reports, from a total of 12.7 million words contained in the entire collection. The place-names were annotated using an XML format described in detail by Grover et al (2010). We transformed the output into a form that could be loaded more easily into CQPweb and removed any mark-up not required for our purposes. This resulted in a series of files containing running text with place-names annotated in place with <enamex> XML tags containing attributes from the gazetteer related to that place-name (spatial metadata). An example marked-up sentence is given below to illustrate this.

> In <enamex sw="w1085" long="0.166493184864521" lat="52.66304588317871" type="ppl" gazref="unlock:9292922" name="Wisbech" conf="1.372404556">Wisbech</enamex> sanitary supervision commenced soon after the cholera epidemic of 1854.

The resulting corpus was then loaded into CQPweb, via part-of-tagging with CLAWS (Garside and Smith, 1997) and semantic annotation with USAS (Rayson et al, 2004), and subsequently into the GCT ready for performing queries, as described in Section 2. Metadata for each text was already present as part of the Histpop digitisation process, including the dates covered by the text and the area covered (e.g. England, Scotland, Great Britain).

The history of the Registrar General's reports starts in 1837. After the Births and Deaths Registration Act of 1836, England and Wales implemented for the very first time the compulsory civil registration of vital events including marriages (Higgs, 2004). The General Register Office (GRO) was in charge of

collecting this information per individual, which was provided by local networks of registrars assigned to the diverse registration districts. From these documents, reports were produced and the information was analysed by the Registrar General and Statistical Superintendents that were usually former qualified Medical Officers of Health. The participation of medics as registration officers did not only help to improve the methods of registration, but also to expand the focus of the reports to relevant issues like public health and sanitation. The influence of these reports was reflected in actions such as the creation of the General Board of Health after the 1848 cholera outbreak, and the governmental interest shown in the acquisition of extensive information about different diseases such as cholera, diarrhoea, dysentery, measles and small-pox among others and their impact in the population. One of the reasons for this interest was that these illnesses were among the most common causes of death during the 19th century, and were also the focus of most public attention due their high mortality and contagion. The cholera epidemics experienced in 1848-49 and 1854 in places like London and the mortality records consequently gathered by the GRO, played an essential role in the research carried out by John Snow and Henry Whitehead, leading eventually to the discovery of the linkage between contaminated water and the disease (Cameron and Jones, 1983; Eyler, 2001; Vinten-Johansen et al, 2003). Thus the Registrar General's reports were crucially important source which led to major improvements in living conditions in industrial towns and cities, the development of public health policies, and the allocation of resources for fighting disease and improving life chances.

The statistical (quantitative) information on health and mortality contained in these reports has been explored historically and geographically on numerous occasions (i.e. Razzell, 1965; Woods and Woodward, 1984; Woods, Watterson and Woodward, 1988; Woods, Watterson and Woodward, 1989; Hardy, 1994; Williams and Galley, 1995; Woods and Shelton, 1997; Laxton and Wiilliams, 1989; Woods, 1985; 2000; 2005; 2007; Gregory, 2008). Far less attention has been given to the reports as (qualitative) historical documents (Szreter, 1991a; 1991b; Lewes, 1991; Goldman, 1991; Higgs, 1991; 1996; Mooney, 1997; Woods, 2007). Part of the explanation for this is that for several

decades computers have been able to handle the large volumes of statistical information that the reports contain, however, they have not been able to handle the very large quantities of text that accompany them.

Looking to address this issue, it is our hypothesis that key questions such as the geographies by which particular diseases were documented and discussed in the reports can be more fruitfully explored through a combination of linguistic and spatial approaches. These approaches furnish new insights into the spatial and temporal patterns underlying the Registrar General's representation of particular important events and enable us to explore whether there were geographical biases in these representations. With this in mind, we investigated whether the combination of NLP and spatial methodologies suggested here may help historians to identify patterns within historical texts, and thus generate new pathways for research in the history of civil registration and the public health movement.

The paper explores the geographical distribution of a set of well-known diseases (Cholera, Diarrhoea and Dysentery - ChDiDy) as depicted in the texts of the reports produced by the GRO during the decades from the 1840s to the 1880s in England and Wales [1]. These diseases were chosen due to the extensive documentation available about them for this period, which allow us to assess the results from the methods implemented. This involved several steps (Fig. 2). In the first place, using the GCT , we carried out a search for all instances of the words "Cholera", "Diarrhoea" or "Dysentery", and looked for place-name mentions within each sentence .This resulted in a geo-database which we then used to investigate how these places were related to the diseases in the reports. More specifically, we aimed to define whether the analysis of the distribution of references to diseases in the texts for different decades could point up unexpected patterns or important events portrayed in the reports, identifying not only particular places associated to these diseases and changes in their

[1] These diseases were analysed together due to the fact that in the early reports produced by the GRO, Cholera, Diarrhoea and Dysentery were grouped.

depiction over time, but also determining whether some of these references clustered in particular regions and the possible reasons behind this.

4. Spatial analysis of the corpus

The first step in exploring the corpus from the spatial perspective was to create simple point data from all place-name instances related to ChDiDy per decade. Although these distribution maps can provide a general sense of the areas included by the reports during the decades covered in this experiment, they do not provide any further element relevant for interpretation. Mapping the frequencies in which places were associated to ChDiDy may provide an idea of the geographies that were subject of more attention by the Registrar General for any given reason (Fig. 4). Although looking at these figures it may become apparent that main urban areas and ports such as London, Manchester, Newcastle and Liverpool amongst others, were always related to these disease words in the reports, these maps do not clarify whether these references occur randomly within the overall distribution of place-names in the corpus, and whether they tend to group or aggregate in certain locations.

In order to test this, a global version of the ***Besag-Newell*** method was applied to examine whether there was clustering of references of these diseases in any area of the study region. This adaptation of the method is available in Clusterseer and is not concerned to find where possible significant clusters are, but whether clustering exists. This helps to indicate if there is a spatial pattern that is unlikely to have occurred by chance. While the original method (Besag and Newell, 1991) compares number of disease cases with population at risk, in this case, the number of disease collocates for each decade was compared with the overall distribution of place-names in order to answer the research question:

a. Is there clustering of place-names within the Register General reports of each decade associated to cholera, diarrhoea and dysentery (ChDiDy) in any area of England and Wales?

The method as implemented in Clusterseer consisted in 'scanning' the data for each decade by district, looking to find what appear as unusual clusters of disease collocates calculating *(l)* local and *(r)* global statistics. This is accomplished by centring a circular window in each region. The window is then expanded to include all neighbouring regions until the total number of disease collocates within reaches a threshold *(k)* (Durbeck et al, 2002). This threshold represents the given size of a cluster and it was specified by us. The number of regions that are necessary in order for the window to contain *k* is known as the local statistic *(l)*. The disease collocate count inside the window is then compared to that of the place-name instances happening in that region, looking to check whether the number of cases in the former is unlikely for the number in the latter. The idea behind this is to test the null hypothesis *(H_o)* that states that the observed total number of collocates is distributed at random amongst the place-names, this is to say, that there is no clustering in the disease collocates. In this manner, the place-name instances inside the window should be proportional to the count of collocations. If this is not the case, the null hypothesis can be rejected. This is defined by the P-values obtained in this test that have to fall below the significance level set in this case to 0.05 in order to be significant. Finally, the Global statistic *(r)* which constitutes the total number of significant local clusters is evaluated through Monte Carlo simulations (999 runs in this case).

One of the disadvantages of this method is that the definition of (k) or the number of cases that form a cluster is decided by the researcher. This is of importance because, as pointed out by Waller and Turnbull (1993), the significance of *(l)* will depend on the chosen value of (k), and this may have considerable implications. An adequate cluster size might be easier to ascertain in more traditional research such as epidemiology, where the reporting of actual occurrences of diseases may allow investigating and setting up a satisfactory threshold. In the case of these historical texts, there is no objective way to establish an appropriate cluster size of disease collocations within the corpus.

Therefore, diverse tests with different thresholds of collocations as (k) were carried out and a test of multiple comparisons was done in order to assess these results.

Once it is established that some clustering existed, the ***Kulldorff's spatial scan method*** was used to identify the location of such clusters per decade. In this case, the analysis was implemented using the SaTScan program to answer the research question:

b. Where are the references to Cholera, Diarrhoea and Dysentery significantly elevated?

This method also implements a circular window to scan the entire study region (Kulldorff, 1997). The window varies in size from the smallest distance to a specified limit, which in this case covered 50% of the whole study region. The method creates a number of distinct circles with different sets of neighbouring data location within them. In this case, places with ChDiDy collocates were set as cases while place-name instances in general (excluding those that collocate with ChDiDy) were set as controls. Under the null hypothesis *(H_o)* of no clustering the behaviour of the data should be the same throughout the study region. For each window, a likelihood ratio test statistic is carried out computing if the observed number of cases is unlikely for a window of that size; this is to say, comparing whether the cases inside the scan window are greater than the outside. The statistical significance is assessed through a Bernoulli model and the distribution under the null hypothesis and p-value are established running Monte Carlo simulations. The window associated with the maximum likelihood ratio is defined as the primary cluster candidate occurring not by random chance. Secondary clusters are also reported and they are defined as statistically significant when their log likelihood ratio (LLR) is greater than the critical value established for the significance level desired, which in this case was also 0.05. One of the advantages of this method is that every window could be a potential cluster, so its size does not need to be specified beforehand.

The last objective of this study was to explore how similar the spatial distributions of ChDiDy collocations were in the studied decades, allowing us to identify the geographies most associated to

the diseases in comparison to other periods and significant changes in these associations over time. In this case the ChDiDy collocates for the decade of 1840 was used as baseline compared to the other decades using the ***Similarity Index test*** to answer the research question:

c. How do the references to these diseases change geographically over time? Do the concentrations of references identified, cluster in the same places?

The method, as implemented in the application SpatialTest, measures at a local level the similarity between two different spatial point patterns calculating for each area a similarity index. The S-Index represents the proportion of spatial units that have a similar spatial pattern in both distributions, ranging from 0=no similarity to 1=perfect similarity. In order to compare both datasets, one is defined as base and the other as test or reference. From the test dataset, 85% of the points are randomly sampled and then aggregated by district. In order to assess statistical significance, random sampling is repeated 200 times. The aggregated counts by region are transformed to percentages and a nonparametric 95% confidence interval for each spatial unit is created (Andresen, 2009; Andresen and Malleson, 2011). If the base percentage value of a spatial unit falls within the confidence interval calculated for that same unit both point patterns are considered as similar. Finally, the similarity index is calculated and the results from the Monte Carlo simulations can be mapped, revealing where the spatial pattern is significantly different (Andresen and Malleson, 2011).

5. Results

The results from the collocation analysis show the incidence in which Cholera, Diarrhoea and Dysentery were associated to places and covered by the GRO during each decade. Plotting the frequency of these associations, it can be observed that a large number of these references occurred during the 1840s, having a significant increase that reached a peak during the 1860s, to almost disappear towards the 1880s (Fig. 5). The P-values obtained by the global Besag-Newell method

applied, allowed us to reject the null hypotheses tested at a significance level of 0.05, establishing that ChDiDy collocations tend to cluster in particular places (Table 1).

The results from the Kulldorff spatial scan method allowed the identification of a primary cluster in each decade and according with the log likelihood ratios (LLR) obtained, further statistically significant clusters where also located (Table 2). For all decades, the primary cluster observed corresponded to London and its surrounding areas, with the exception of 1870s, for which the primary cluster was recorded in the area of Newcastle. As will be discussed below, the results from these analyses were mapped per decade and the clusters were symbolised by cluster rank according to their LLR.

To explore changes over time in the distribution of disease collocations, four tests were carried out. Indices of similarity were calculated for 1840-1850, 1850-1860, 1860-1870 and 1840-1870 (Table 3). The results were used to explore whether the spatial patterns of references of these diseases changed over time and how stable they were between decades.

6. Discussion

Comparing the references to cholera, diarrhoea and dysentery related to places with the mortality rate of these diseases, it can be observed that they follow a similar trend. From 1850 onwards similarly to ChDiDy collocations, deaths from these diseases reached a peak during the 1860s decreasing also eventually towards the 1880s (Fig. 6).

From this observation it does not follow that there is a correlation between them, or that the RG mentioned the diseases in direct proportion to the amount of deaths caused by them. What is shown is that the frequency in which the diseases were mentioned in the reports and its relationship to places reveal important events happening and discussed by the RG during this period, and therefore it provides interesting insights to the actual reports. This becomes more evident plotting

the ChDiDy collocations per year, where four main peaks in the frequencies can be identified for 1849, 1854, 1866 and 1868 (Fig. 7).

By the beginning of the 1840s the debate over the link between disease and living conditions was very alive. As evidenced by Chadwick's report published in 1842 (Chadwick, 1842), public health was increasingly becoming a topic of attention. Nevertheless, understanding of the mechanisms of disease transmission and the role that certain hygienic conditions played mainly on the cities was still poor. In October of 1848 the second largest outbreak of cholera struck England. Between this year and 1849, more than 53,000 people died from this disease (Snow, 2002). This event would become of significant importance in the history of epidemiology as it enabled the physician John Snow to collect part of the evidence that would lead him to formulate his theory of causation and transmission of the disease published on this work 'On the mode of Communication of Cholera' in 1855. The 1849 epidemic is reflected on the first peak in the graphic produced from our collocation analysis, which depicts how a large amount of places were related to cholera, diarrhoea and dysentery in the corpus during this year (Fig. 7). This relationship is particularly outstanding in the case of London where the Kulldorff method revealed it as the primary cluster scoring a LLR of 124.889, and exceeding by far other secondary clusters identified (Fig. 8). It is interesting to note that despite the total mortality in this city being high (14,137 deaths), it was not among the locations with the highest mortality rate, proportional to its population, during the epidemic (Farr, 1868: [C.4072]). The identification of London as the primary cluster suggests that the reports paid significantly more attention to London than elsewhere. This is probably due to the fact that during this cholera epidemic, the capital had the largest total number of deaths in the country. Places that experienced the highest mortality rate were also identified in the analysis. Methyr Tydfil, for example, was at that time the most important industrial town in Wales and it was also the most afflicted district in the country, presenting the highest mortality in proportion to population with 251 deaths per 10,000 habitants (Farr, 1868: [C.4072] 18). Along with Methyr Tydfil, other places like Sculcoates and Salisbury were recognised by the analysis as part of the secondary clusters with a LLR

of 44.178, and they figure in the reports as second and third place respectively in terms of highest mortality in all the country (Farr, 1868: [C.4072] 18). The other urban centres that were most affected by this epidemic in the GRO reports were Liverpool, West Derby, Tynemouth, Hunslet, Leeds, Newcastle, Cardiff, Portsea, Bristol, Southampton and Newport. These centres were also recognised by the spatial analysis as the sites of statistically significant clusters (Table. 2).

The next peak revealed in Figure 7 by the collocation analysis corresponds to another key date in the history of disease and epidemiology in England. Although the involvement of medical experts in policy making increased and the progress of the sanitary reforms was advancing, a further epidemic took place in 1854. By the end of this year, more than 40,000 died from cholera and diarrhoea alone in all the country (Farr, 1868: [C.4072] 1). In London, the area of the St. James district (Soho) was particularly affected and it was this event that helped the research by John Snow and Rev. Henry Whitehead to finally establish the waterborne character of the disease. While Whitehead helped John Snow to access and gather crucial information about the cases emerging in the community of Soho, the Registrar General's reports provided by William Farr played also a fundamental role in his investigation. The collocation analysis show that the RG documented this epidemic extensively and the cluster analysis for this decade identified five statistically significant clusters pointing out St. James and other places surrounding London as part of the primary cluster with a LLR of 33.84 (Fig. 9). The identification of London and its surroundings as the primary cluster once again seems to indicate that the RG focused on the city that suffered the greatest number of deaths during this epidemic (10,738 deaths). As in the previous case, places with the highest mortality rates, such as Milton, Towcester, and Brenford (Farr, 1868: [C.4072] 18), were also identified with the highest LLR after London (Fig. 9).

Other places recognised as part of the secondary clusters are related to further particular disease events. For instance, the substantial mentions of West Ham identified within the third statistically significant cluster for this decade, is related to the peak of 1857 in the graphic of the collocation

analysis. This corresponds to a small cholera outbreak registered in the place, which was further investigated by John Snow (Snow, 1857).

Other minor peaks represented in the collocation graphic (Fig. 7) such as the one recorded for 1860, and the identification in the cluster analysis of places such as Hull in the north-west of England, points out the discussions on the reports about the concerns raised regarding the vulnerability to these diseases in particular places, where living conditions were difficult and other factors such as how the presence of travellers could play an important role in the 'transportation' of the disease (5th cluster- Fig. 10). The next small peak on the collocation graphic coincides with the discussion of the 1863 global cholera pandemic that, despite of not affecting Britain in that particular year, was a concern among the general public, particularly in places that served as major ports such as Liverpool, Southampton, Newcastle, and London (Farr, 1868: [C.4072] xii). These discussions and concerns are reflected on both, the linguistic and spatial analyses. As in the previous cases, the collocation analysis identified the last important cholera event discussed in the reports, showing a large number of collocations for 1866. Along with Liverpool, Newcastle and London, the cluster method identified again the places reported with the highest mortality rates during this cholera epidemic such as Swansea, West Derby, Methyr Tydfil and Birkenhead (Farr, 1868: [C.4072] 18). Due to the reforms already implemented in the sewage system and a better understanding of the disease, the epidemic of 1866 was comparatively small and it has attracted historically less attention. Nevertheless, there were more than 31,500 deaths from cholera and diarrhoea in the country, from which more than 8,500 happened in London alone (Farr, 1868: [C.4072] 2). In this manner, the importance of the outbreak is still reflected on the frequency of disease references detected on the corpus, as well as on the spatial analysis which identified London as the primary cluster once more.

It is interesting that the largest peak detected (1868) by the collocation analysis does not correspond directly to a disease outbreak (Fig. 7). It is in fact product of a particular chapter of the history of the GRO. Due to the impact of previous epidemics, the RG decided to write a dedicated report as

supplement exclusively concerned with these diseases. This document formed part of a singular section of the registrar general reports. Written by William Farr, and entitled 'Report on the Cholera Epidemic of 1866 in England', this document was appended to the twenty-ninth annual report of the Registrar General in 1868, and is thus part of the corpus analysed in this study. In this document Farr offers a detailed account on the diverse cholera epidemics beginning with the first outbreak of 1831. He also discusses the different theories of its spread and contagion and provides comprehensive information regarding the diverse places in which cholera and diarrhoea took particular relevance in England. In this manner, the large peak observed in our graphic highlights the presence of this special report within the corpus. The analysis effectively identified an unusual number of references to these diseases in 1868, year in which his volume was published. The key point is thus that by this stage the Registrar General was in a position to publish extensive research on cholera epidemics in anticipation of a new epidemic arriving, however, lessons learned from previous epidemics and actions taken as a result meant that this global epidemic never really took hold in Britain.

Finally, for the 1870s there was a significant decrease on the mention of these diseases, especially cholera. Only two minor peaks in the collocation frequency were recognised for 1871 and 1873. This can be attributed to the considerable improvement of sanitary conditions and increased awareness about the mode of transmission of Cholera. The progress made especially in London, is reflected in the fact that the number of deaths from this disease dramatically decreased during this decade and there was no other epidemic after 1866 (General Register Office, 1800: [C.2568] 231), but also in the historical reports where the references to the disease associated with places become scarcer and their spatial distribution changes. This is shown in the cluster analysis where the attention of the reports, in terms of geographies linked to the diseases, can be seen to shift in this decade towards Newcastle (Fig. 11). This city was determined this time as the primary cluster, while London fell to the secondary clusters. Although Newcastle was among the places most affected in previous epidemics (mainly the occurred during 1831-31), the number of deaths for the 1870s remained small and the disease never affected the region as experienced before. Nevertheless, as shown by the

cluster analysis the discussion carried out in relation to ChDiDy still focused on the major and industrial cities such as Liverpool, Manchester, Nottingham and London.

The results of the spatial point pattern test analysing changes over time in the distributions of references to cholera, diarrhoea and dysentery are shown in Table 3 which shows the similarity index observed between decades. As said before, a value of 1 represents an equal spatial pattern, while 0 represents a completely different one. As seen in the table, the spatial pattern of these diseases was moderately variable over time and it was reasonably close to perfect similarity especially in the comparison between the distributions for the 1840s and 1870s. It is interesting to note that the index values are relatively stable over time meaning that the distribution of the mentions of ChDiDy concentrated regularly in the same places. This is unsurprising as it is probably due to the fact that particular locations were relatively more vulnerable to these diseases such as cities and major ports, and therefore, the mention of these diseases changed only subtly over time.

Finally, is worth observing that for three consecutive decades (1840-1860), the capital was the place most frequently mentioned in association to these diseases despite of not having the highest mortality rates during the major outbreaks in the country. The historical reasons for this are probably varied and there is still the need to explore the possible causes that lead the registrar general to focus in the capital, clarifying whether the reports paid more attention to urban areas. Nevertheless, an important outcome in this particular case is that through the development of these techniques is possible to finally address this kind of questions, allowing us to explore historical texts in an innovative way.

7. Conclusions

The combination of spatial and linguistic methodologies deployed in this paper for the exploration of historical texts is new. The main goal behind this first case study was to show how the combination of language analysis and spatial methods can help in the analysis of large corpora with a geographic

nature. Although there is still work to be done in the implementation of these approaches, this paper shows that combining Natural Language Processing and Corpus Linguistics techniques with GIS-based spatial analysis can provide new insights into the geographies within large volumes of text.

In this particular case, the collocation analysis made it possible to identify in an automated way the most relevant episodes in the history of cholera portrayed in the text of the Registrar General's reports. In addition to that, it highlighted the presence within the digital collection of one of the most noteworthy documents related to the history of these diseases, the 1868 Report on the cholera epidemics written by Farr. The spatial analyses were able to reveal, not only the places related to these events, but also the role they played and the importance they had in such events as described in the historical sources. In this manner, our research shows the potential that linguistic and spatial techniques have for the analysis of large volumes of texts and to compare them with evidence from both other texts and quantitative sources. The results from these techniques highlighted the parts in a corpus comprised of more than 200,000 pages that were intrinsically related to our topic of interest, allowing us to identify effectively and in an efficient way areas in the corpus that called for close reading. These results enable us also to think further not only about the geographies related to this topic, but also to analyse the consideration paid to them in the historical reports. In addition to that, the methodologies combined and employed here will allow us to examine the mention of diseases (among many other topics) not only in official reports, but also in conjunction to other large textual collections such as newspapers. This will facilitate a comparison of the importance given and perceptions that official and non-official sources had of diseases during the nineteenth century, and the social impact provoked and portrayed in different kinds of historical sources.

The semi-automated identification of the most significant events, the places related to them and the spatial patterns portrayed in a large corpus open new exciting possibilities not only in History as a

discipline but also in other areas dealing with texts of a geographical nature such as Archaeology or Literary Studies. From the results of this pilot study, we believe that these approaches can be implemented to explore large corpora, detecting what places the corpus is talking about; what is being said regarding these places, and the changes those places experienced over time in their representation within the corpus.

Acknowledgements

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant "Spatial Humanities: Texts, GIS, places" (agreement number 283850).

Bibliography

Andresen, Martin A. 2009. Testing for similarity in area-based spatial patterns: a nonparametric Monte Carlo approach. *Applied Geography* 29(3): 333–345.

Andresen, Martin A, and Nicolas Malleson. 2011. Testing the stability of crime patterns: Implications for theory and policy. *Journal of Research in Crime and Delinquency* 48(1): 58–82.

Anselin, Luc. 1995. Local indicators of spatial association—LISA. *Geographical analysis* 27(2): 93–115.

Besag, Julian, and James Newell. 1991. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*: 143–155.

Brooker, Simon, Siân Clarke, Joseph Kiambo Njagi, Sarah Polack, Benbolt Mugo, Benson Estambale, Eric Muchiri, Pascal Magnussen, and Jonathan Cox. 2004. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. *Tropical medicine & international health* 9(7): 757–766.

Cameron, Donald, and Ian G Jones. 1983. John Snow, the Broad Street pump and modern epidemiology. *International journal of epidemiology* 12(4): 393–396.

Chadwick, Edwin. 1842. 1 *Report on the sanitary condition of the labouring population og Great: Britain: supplementary report on the results of special inquiry into the practice of interment in towns*. HMSO.

Chen, Jin, Robert E Roth, Adam T Naito, Eugene J Lengerich, and Alan M MacEachren. 2008. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of US cervical cancer mortality. *International journal of health geographics* 7(1): 57.

Chrysanthi, Angeliki, Patricia Murrieta-Flores, and Constantinos Papadopoulos. 2012. 2344 *Thinking Beyond the Tool. Archaeological Computing and the Interpretive Process*. Archaeopress.

Cooper, David, and Ian N Gregory. 2011. Mapping the English Lake District: A literary GIS. *Transactions of the Institute of British Geographers* 36(1): 89–108.

d'Aignaux, Jérôme Huillard, Simon N Cousens, Nicole Delasnerie-Lauprêtre, Jean-Philippe Brandel, Dominique Salomon, Jean-Louis Laplanche, Jean-Jacques Hauw, and Annick Alpérovitch. 2002. Analysis of the geographical distribution of sporadic Creutzfeldt-Jakob disease in France between 1992 and 1998. *International journal of epidemiology* 31(2): 490–495.

Darby, H Clifford. 1953. On the relations of geography and history. *Transactions and Papers (Institute of British Geographers)* (19): 1–11.

Durbeck, H, D Greiling, L Estberg, A Long, and G Jacquez. 2002. ClusterSeer: Software for Identifying Event Clusters User's Guide. TerraSeer.

Eyler, John M. 2001. The changing assessments of John Snow's and William Farr's cholera studies. *Sozial-und Präventivmedizin* 46(4): 225–232.

Farr, William. 1868. *Report on the cholera epidemic of 1866 in England: supplement to the twenty-ninth annual report of the registrar-general of births, deaths, and marriages in England*. George E. Eyre and William Spottiswoode.

Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102-121.

Gilbert, Pamela K. 2008. *Cholera and Nation: Doctoring the Social Body in Victorian England*. SUNY Press.

Goldman, Lawrence. 1991. Statistics and the science of society in early Victorian Britain; an intellectual context for the General Register Office. *Social History of Medicine* 4(3): 415–434.

Green, Chris, Robert D Hoppa, T Kue Young, and J F Blanchard. 2003. Geographic analysis of diabetes prevalence in an urban area. *Social science & medicine* 57(3): 551–560.

Gregory, I, and Humphrey Southall. 2000. Spatial frameworks for historical censuses: the Great Britain Historical GIS.

Gregory, Ian N. 2003. *A place in history: A guide to using GIS in historical research*. Oxbow Oxford.

———. 2008. Different places, different stories: Infant mortality decline in England and Wales, 1851--1911. *Annals of the Association of American Geographers* 98(4): 773–794.

Gregory, Ian N, and Andrew Hardie. 2011. Visual GISting: bringing together corpus linguistics and Geographical Information Systems. *Literary and linguistic computing* 26(3): 297–314.

Gregory, Ian N, Jordi Marti-Henneberg, and Francisco J Tapiador. 2010. Modelling long-term pan-European population change from 1870 to 2000 by using geographical information systems. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173(1): 31–50.

Grover, Claire, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368(1925): 3875–3889.

Hardie, Andrew. 2012. CQPweb-combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380–409.

Hardie, Andrew, Tony McEnery, and Scott Piao. 2010. Historical Text Mining and Corpus-Based Approaches to the Newsbooks of the Commonwealth.

Hardy, Anne. 1994. 'Death is the cure of all diseases': Using the General Register Office cause of death statistics for 1837--1920. *Social History of Medicine* 7(3): 472–492.

Higgs, Edward. 1991. Disease, febrile poisons, and statistics: the census as a medical survey, 1841--1911. *Social History of Medicine* 4(3): 465–478.

———. 1996. The Statistical Big Bang of 1911: Ideology, Technological Innovation and the Production of Medical Statistics. *Social history of medicine* 9(3): 409–426.

———. 2004. *Life, death and statistics: civil registration, censuses and the work of the General Register Office, 1836-1952*. University Of Hertfordshire Press.

Kulldorff, Martin. 1997. A spatial scan statistic. *Communications in Statistics-Theory and methods* 26(6): 1481–1496.

Laxton, Paul, and Naomi Williams. 1989. Urbanization and infant mortality in England: A long term perspective and review. *Urbanisation and the Epidemiologic Transition. Uppsala, Sweden: Uppsala University*: 109–135.

Lewes, Fred. 1991. The GRO and the Provinces in the Nineteenth Century. *Social History of Medicine* 4(3): 479–496.

Llobera, Marcos, and Tim J Sluckin. 2007. Zigzagging: Theoretical insights on climbing strategies. *Journal of theoretical biology* 249(2): 206–217.

Lock, Gary R. 2000. *Beyond the map: Archaeology and spatial technologies*. IOS Press.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus linguistics: method, theory and practice*. Cambridge University Press.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Mooney, Graham. 1997. Professionalization in public health and the measurement of sanitary progress in nineteenth-century England and Wales. *Social history of medicine* 10(1): 53–78.

Moretti, Franco. 1999. *Atlas of the European Novel: 1800-1900*. Verso.

———. 2005. *Graphs, Maps, Trees: Abstract models for a literary history*. Verso.

———. 2013. *Distant Reading*. Verso.

Murrieta-Flores, Patricia, Christopher Donaldson, C.J. Rupp, David Cooper and Ian Gregory. 2013. Digital Literary Geographies: A Spatial Analysis of Lake District Landscape Writings. Poster presented at *GIS Research UK (GISRUK)*, University of Liverpool, U.K.

Murrieta-Flores, Patricia, David Cooper, and Ian Gregory. 2012. Spatial Humanities: Exploring and Analysing Texts within a GIS Environment. Paper presented at the *Digital Humanities Congress*, University of Sheffield, U.K.

Rayson, P., Archer, D., Piao, S. L. & McEnery, T. 2004. "The UCREL semantic analysis system". In Proceedings of the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks, in Association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, 7–12.

Rayson, Paul, Alistair Baron, and Andrew Hardie. 2012. Which 'Lancaster'do you mean? Disambiguation challenges in extracting place names for Spatial Humanities. Paper presented at the *Digital Humanities Congress*, University of Sheffield, U.K.

Razzell, Peter E. 1965. Population Change in Eighteenth-Century England. A Reinterpretation. *The Economic History Review* 18(2): 312–332.

Register General Office. 1800. BPP 1880 XVI [C.2568] 231 *Forty-first Annual report of the registrar-general (1878)*.

———. 1868. *Report on the Cholera Epidemic of 1866 in England: Supplement to the Twenty-ninth Annual Report of the Registrar-general of Births, Deaths, and Marriages in England*. George E. Eyre and William Spottiswoode. http://books.google.co.uk/books?id=ftrkPQAACAAJ.

Rupp, C.J., Christopher Donaldson and Patricia Murrieta-Flores. 2013. Integrating Corpus Linguistics and Spatial Technologies for the Analysis of Literature. Paper presented at *Corpus Linguistics 2013*, Lancaster University, U.K.

Rupp, C.J., Paul Rayson, Alistair Baron, Christopher Donaldson, Ian Gregory, Andrew Hardie, Patricia Murrieta-Flores. 2013. Customising geoparsing and georeferencing for historical texts. In proceedings of IEEE Big Humanities workshop, IEEE Big Data conference, San Francisco, October 2013.

Snow, John. 1857. On the Origin of the Recent Outbreak of Cholera at West Ham. *British Medical Journal* 1(45): 934.

Snow, Stephanie J. 2002. Commentary: Sutherland, Snow and water: the transmission of cholera in the nineteenth century. *International journal of epidemiology* 31(5): 908–911.

Szreter, Simon. 1991a. Introduction: the GRO and the historians. *Social History of Medicine* 4(3): 401–414.

———. 1991b. The GRO and the public health movement in Britain, 1837--1914. *Social History of Medicine* 4(3): 435–463.

Thomas, Amanda J. 2010. *The Lambeth Cholera Outbreak of 1848-1849: The Setting, Causes, Course and Aftermath of an Epidemic in London*. McFarland.

Vinten-Johansen, Peter. 2003. *Cholera, chloroform, and the science of medicine: a life of John Snow*. Oxford University Press, USA.

Waller, Lance A, and Bruce W Turnbull. 1993. The effects of scale on tests for disease clustering. *Statistics in Medicine* 12(19-20): 1869–1884.

Warf, Barney, and Santa Arias. 2008. 26 *The spatial turn: Interdisciplinary perspectives*. Routledge.

Wheatley, David, and Mark Gillings. 2004. *Spatial technology and archaeology: the archaeological applications of GIS*. CRC Press.

Williams, Naomi, and Chris Galley. 1995. Urban-rural differentials in infant mortality in Victorian England. *Population Studies* 49(3): 401–420.

Woods, Robert. 1985. The effects of population redistribution on the level of mortality in nineteenth-century England and Wales. *The journal of economic history* 45(03): 645–651.

———. 2000. 35 *The Demography of Victorian England and Wales*. Cambridge University Press.

———. 2005. The measurement of historical trends in fetal mortality in England and Wales. *Population studies* 59(2): 147–162.

———. 2007. Medical and demographic history: Inseparable? *Social history of medicine* 20(3): 483–503.

Woods, Robert I, Patricia A Watterson, and John H Woodward. 1988. The causes of rapid infant mortality decline in England and Wales, 1861--1921 Part I. *Population Studies* 42(3): 343–366.

———. 1989. The causes of rapid infant mortality decline in England and Wales, 1861--1921. Part II. *Population Studies* 43(1): 113–132.

Woods, Robert I, and John Woodward. 1984. *Urban Disease and Mortality: In Nineteenth-Century England*. BT Batsford Limited.

Wright, Gwendolyn. 2005. Cultural History: Europeans, Americans, and the Meanings of Space. *Journal of the Society of Architectural Historians* 64(4): 436–440.

Zhang, Zhijie, Tim E Carpenter, Yue Chen, Allan B Clark, Henry S Lynn, Wenxiang Peng, Yibiao Zhou, Genming Zhao, and Qingwu Jiang. 2008. Identifying high-risk regions for schistosomiasis in Guichi, China: a spatial analysis. *Acta tropica* 107(3): 217–223.

Besag-Newell	Decade	P-Value α=0.05
	1840	0.002
	1850	0.003
	1860	0.006
	1870	0.002

Table 1. Besag-Newell results.

1840's	Clusters	Log likelihood ratio	1850's	Clusters	Log likelihood ratio
α=0.05	1	124.889314	α=0.05	1	33.843408
Critical value	2	44.178283	Critical value	2	16.110283
8.970557	3	38.328177	7.591968	3	14.297706
	4	20.208256		4	13.438797
	5	16.976699		5	11.622168
	6	14.845329			
1860's	Clusters	Log likelihood ratio	1870's	Clusters	Log likelihood ratio
α=0.05	1	918.830208	α=0.05	1	33.899928
Critical value	2	29.520021	Critical value	2	24.056633
9.357238	3	16.417379	8.427948	3	12.725722
	4	14.357114		4	9.516979
	5	13.565095			
	6	12.40873			

Table 2. Kulldorff's spatial scan method results.

Indices of Similarity-Districts				
Cholera, Diarrhoea, Dysentery	1840-1850	1850-1860	1860-1870	1840-1870
	0.73	0.683	0.728	0.8

Table 3. Similarity Index results.

PN_LeftContext	Placename	PN_RightContext
ess of that season . In the South-eastern Counties the deaths were 11,256 in the place of 8400 . The	Epsom	district suffered from scarlatina ; Guild ford from small-pox and measle
lace of 8400 . The Epsom district suffered from scarlatina ; Guild ford from small-pox and measles ;	Farnham	from fever , measles , hooping cough , and diarrhoea . The deaths for th
hooping cough , and diarrhoea . The deaths for the first time exceed the births in Farnham . In the	Bexley	sub-district in Kent there were as many as forty cases of small-pox at on
ry cottage . The South Midland Counties suffered from scarlatina and fever in several districts . In	Oxford	25 deaths occurred from small-pox , and the deaths exceeded the birth
Farnham has for the last seven months been suffering from low fever , pooping-cough , and measles ;	Canterbury	from small-pox , scarlatina , and bronchitis . Alverstoke has suffered se

Fig.1 Collocation example of place-names with the word 'small-pox' as retrieved from the GCT tool.

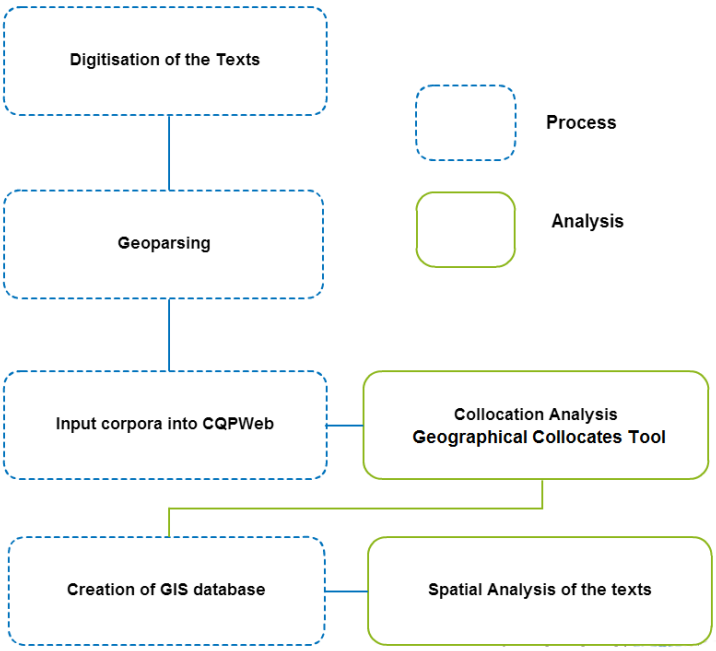

Fig.2 Work-flow of the stages required to analyse the texts.

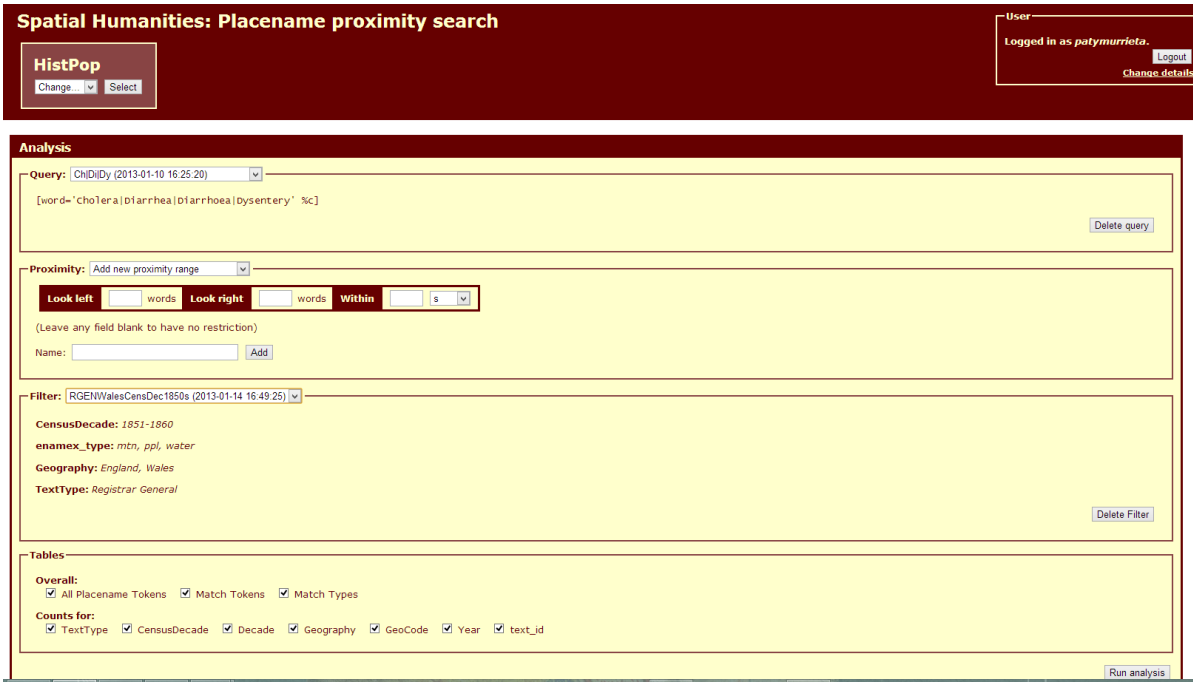

Fig. 3 Screenshot of the Spatial Humanities Geographic Collocates Tool.

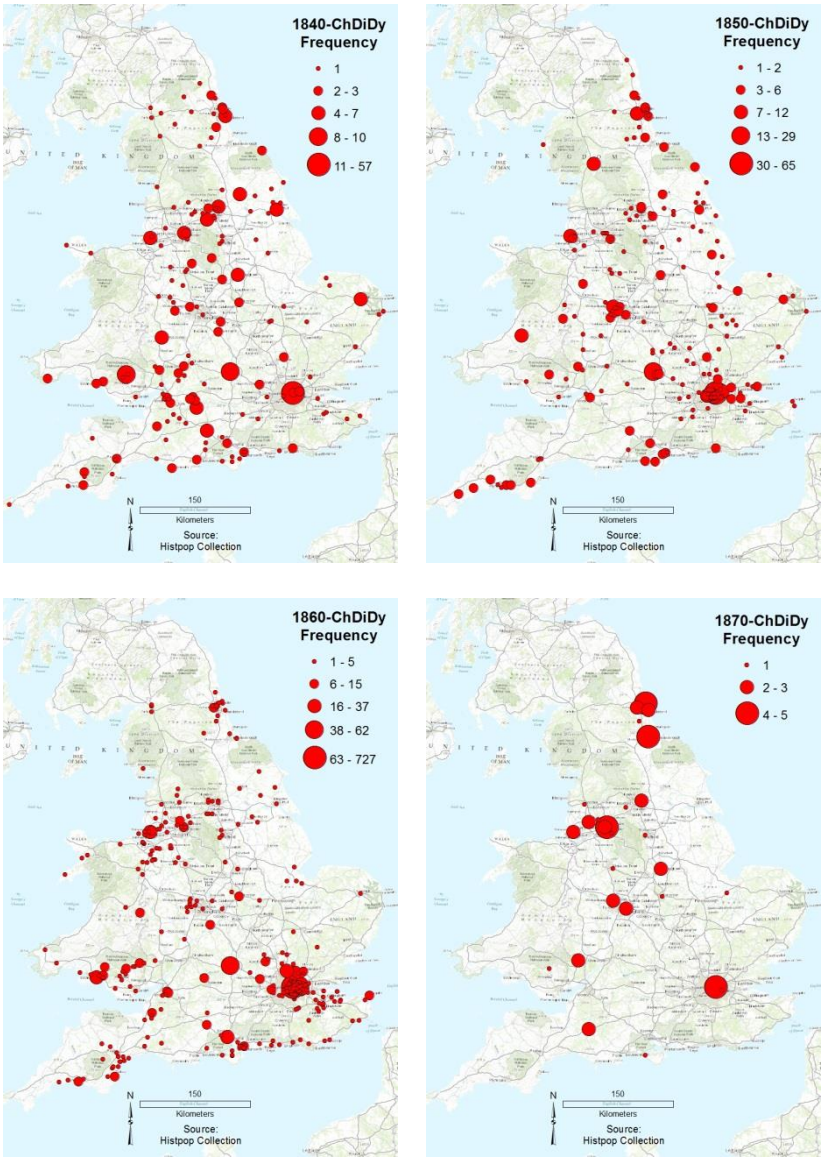

Fig. 4 Maps of the frequency in which places related to ChDiDy were mentioned.

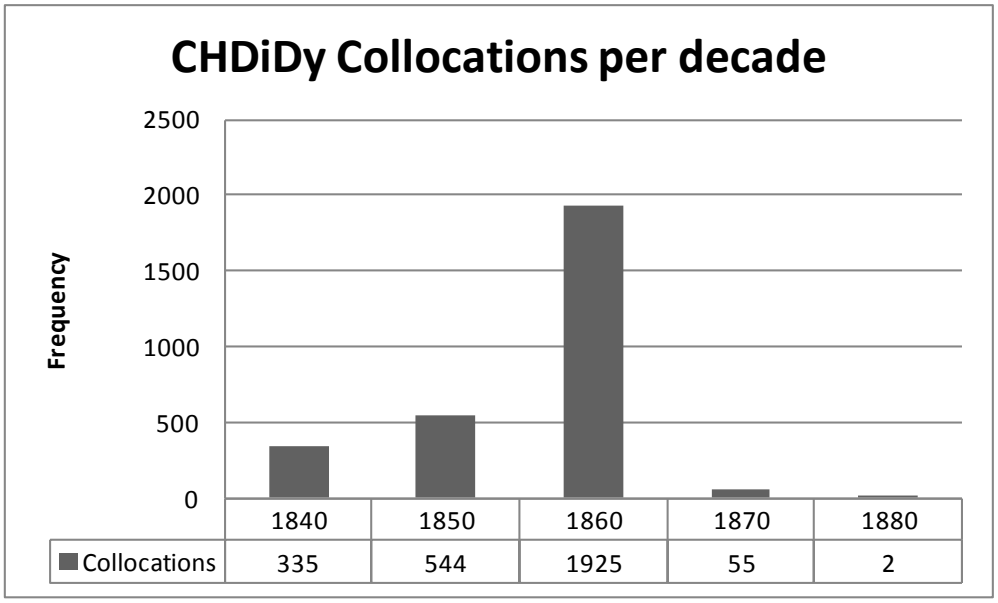

Fig. 5 Frequency of cholera, diarrhoea and dysentery collocations per decade.

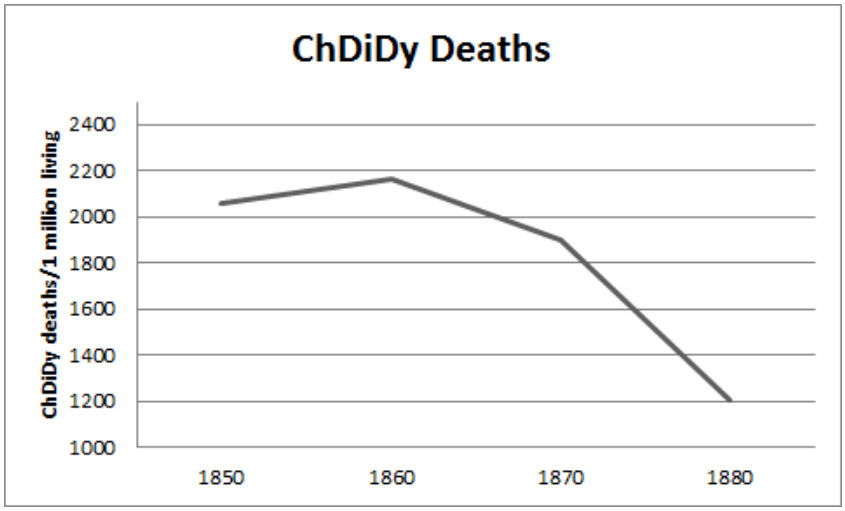

Fig. 6 Deaths from these diseases per decade.

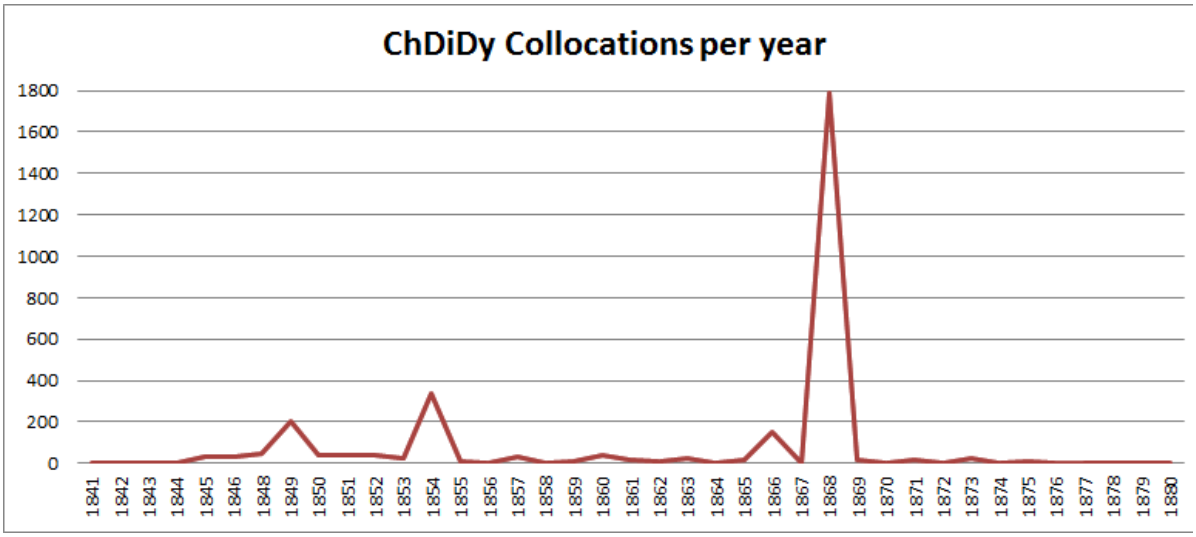

Fig. 7 Frequency of ChDiDy collocations per year.

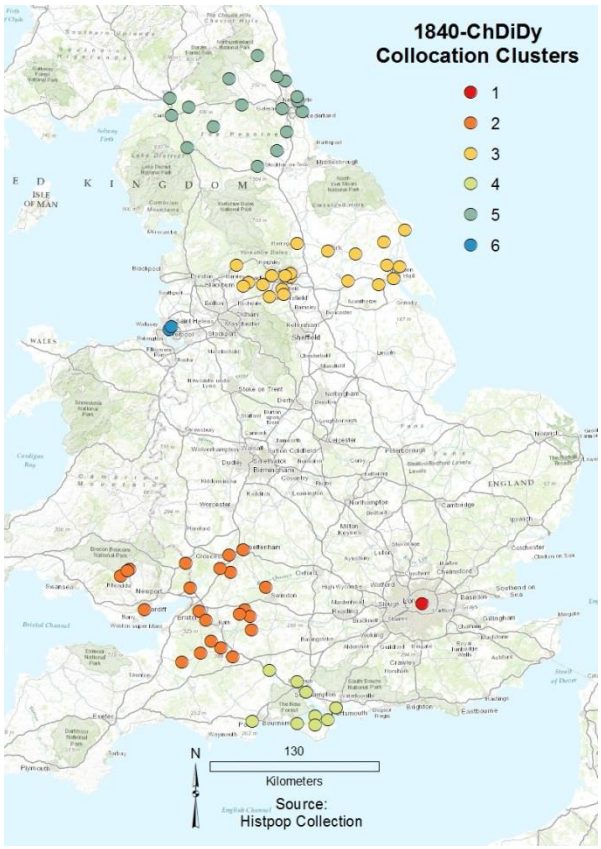

Fig. 8 Clusters identified for the 1840s dataset.

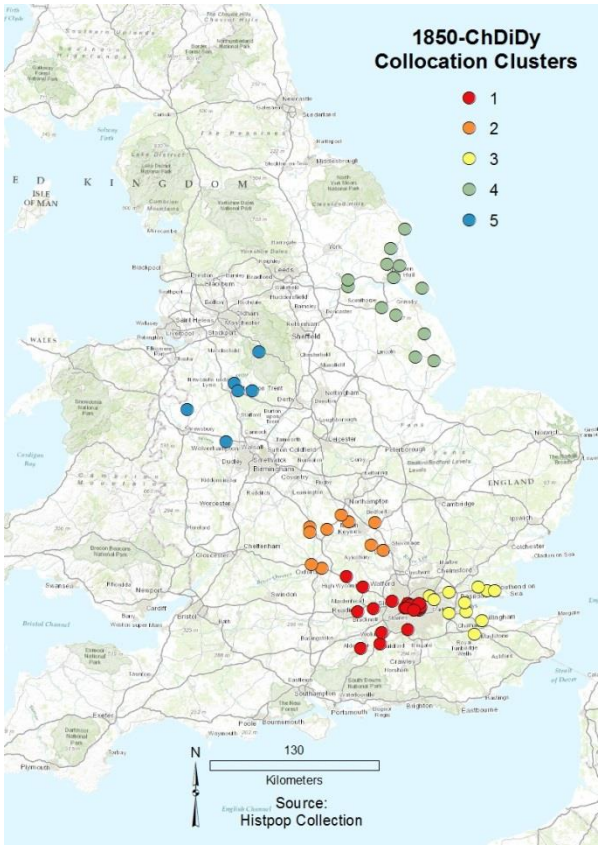

Fig. 9 Clusters identified for the 1850s dataset.

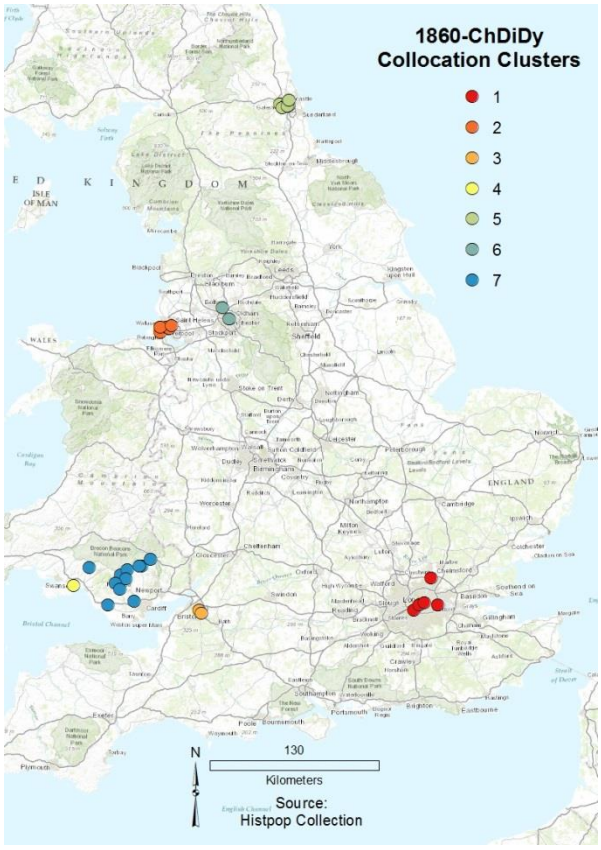

Fig. 10 Clusters identified for the 1860s dataset.

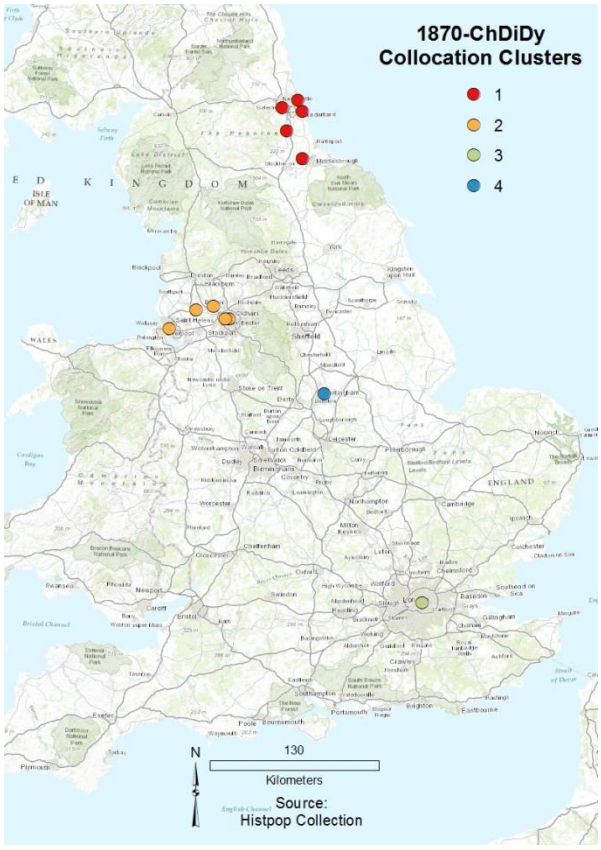

Fig. 11 Clusters identified for the 1870s dataset.